

Multivariate Calibration Standardization across Instruments for the Determination of Glucose by Fourier Transform Near-Infrared Spectrometry

Lin Zhang[†] and Gary W. Small*

Center for Intelligent Chemical Instrumentation, Department of Chemistry and Biochemistry, Clippinger Laboratories, Ohio University, Athens, Ohio 45701

Mark A. Arnold

Optical Science & Technology Center and Department of Chemistry, University of Iowa, Iowa City, Iowa 52242

The transfer of multivariate calibration models is investigated between a primary (A) and two secondary Fourier transform near-infrared (near-IR) spectrometers (B, C). The application studied in this work is the use of bands in the near-IR combination region of 5000–4000 cm^{-1} to determine physiological levels of glucose in a buffered aqueous matrix containing varying levels of alanine, ascorbate, lactate, triacetin, and urea. The three spectrometers are used to measure 80 samples produced through a randomized experimental design that minimizes correlations between the component concentrations and between the concentrations of glucose and water. Direct standardization (DS), piecewise direct standardization (PDS), and guided model reoptimization (GMR) are evaluated for use in transferring partial least-squares calibration models developed with the spectra of 64 samples from the primary instrument to the prediction of glucose concentrations in 16 prediction samples measured with each secondary spectrometer. The three algorithms are evaluated as a function of the number of standardization samples used in transferring the calibration models. Performance criteria for judging the success of the calibration transfer are established as the standard error of prediction (SEP) for internal calibration models built with the spectra of the 64 calibration samples collected with each secondary spectrometer. These SEP values are 1.51 and 1.14 mM for spectrometers B and C, respectively. When calibration standardization is applied, the GMR algorithm is observed to outperform DS and PDS. With spectrometer C, the calibration transfer is highly successful, producing an SEP value of 1.07 mM. However, an SEP of 2.96 mM indicates unsuccessful calibration standardization with spectrometer B. This failure is attributed to differences in the variance structure of the spectra collected with spectrometers A and B. Diagnostic procedures are presented for use with the GMR algorithm that forecasts the successful calibration

transfer with spectrometer C and the unsatisfactory results with spectrometer B.

Coupled with multivariate calibration methods, near-infrared (near-IR) spectroscopy is being given increased attention for use in the determination of clinically relevant analytes such as blood glucose.^{1–5} The principal advantages of near-IR measurements in a clinical setting are the ability to implement a reagentless and nondestructive analysis, the ability to quantify several species with one spectral measurement, and the potential for noninvasive determinations. One practical limitation of this approach is the degradation in the performance of the calibration model that may occur when drift or shift occurs within the spectrometer over time or when predictions are made with spectral data measured with an instrument different from the one used to develop the calibration model. Full recalibration often involves considerable effort and cost.

There are two common multivariate calibration strategies to attack this problem. Methods using the first strategy such as direct standardization⁶ (DS) and piecewise direct standardization⁷ (PDS) transfer the spectra measured under the new situation to be used with the old model by employing a small set of standardization samples. These samples are measured under both old and new conditions, and the resulting spectra are used to derive a transformation model that relates the two sets of spectra.

The second strategy develops calibration models robust to instrumental differences. By employing this approach, we have developed an algorithm to build robust calibration models through selection of calibration samples and weighting of resolution elements using information from a set of standardization samples

- (1) Hall, J. W.; Pollard, A. *Clin. Chem.* **1992**, *38*, 1623–1631.
- (2) Heise, H. M. In *Biosensors in the Body: Continuous In Vivo Monitoring*; Fraser, D. M., Ed.; Wiley: Chichester, U.K., 1997; pp 79–116.
- (3) Heise, H. M. In *Near-Infrared Spectroscopy: Principles, Instruments, Applications*; Siesler, H. W., Ozaki, Y., Kawata, S., Heise, H. M., Eds.; Wiley-VCH: Weinheim, 2002; pp 289–333.
- (4) Fischbacher, C.; Jagemann, K.-U.; Danzer, K.; Muller, U. A.; Papenkordt, L.; Schuler, J. *Fresenius J. Anal. Chem.* **1997**, *359*, 78–82.
- (5) Burmeister, J. J.; Arnold, M. A.; Small, G. W. *Diabetes Technol. Ther.* **2000**, *2*, 5–15.
- (6) Wang, Y.; Veltkamp, D. J.; Kowalski, B. R. *Anal. Chem.* **1991**, *63*, 2750–2756.
- (7) Wang, Z.; Dean, T.; Kowalski, B. R. *Anal. Chem.* **1995**, *67*, 2379–2385.

* To whom correspondence should be addressed. E-mail: small@ohio.edu.

[†] Present address: Pfizer Global R&D, Groton Laboratories, MS 4137, 558 Eastern Point Rd., Groton, CT 06340-5196.

measured under the new conditions.⁸ We have termed this method the guided model reoptimization (GMR) algorithm. One advantage of the GMR method over DS and PDS is that identical samples need not be measured in both situations.

In our initial study,⁸ calibration standardization within the same instrument over a period of time was performed by use of the GMR algorithm. In the work reported here, we further investigate calibration standardization across three different instruments in the context of the determination of physiological levels of glucose in a six-component aqueous matrix by Fourier transform near-IR spectrometry. The GMR, DS, and PDS methods are compared for their effectiveness in performing calibration standardization in this chemical system.

EXPERIMENTAL SECTION

Data Set Design. A set of 80 samples was generated by a randomized experimental design method. The sample matrix consisted of glucose, sodium lactate, urea, sodium ascorbate, alanine, and triacetin in phosphate buffer. The concentration of each component for each sample was randomly assigned across the 80 samples. The concentration ranges for glucose, lactate, urea, ascorbate, alanine, and triacetin were 0.68–34.27, 1.18–33.34, 1.20–30.23, 1.18–22.55, 1.75–38.32, and 0.46–27.95 mM, respectively. Correlation coefficients computed between the component concentrations ranged from –0.18 to 0.11, confirming that each species was independent.

Apparatus and Reagents. Each of the 80 samples was measured with a primary and two secondary spectrometers. All instruments were commercial Fourier transform (FT) spectrometers based on Michelson interferometers configured for the near-IR region. Each instrument was equipped with a tungsten-halogen source, CaF₂ beam splitter, and cryogenically cooled InSb detector. Manufacturers and models are withheld here to preclude unnecessary inferences regarding relative instrument quality. The primary instrument was a one-year-old spectrometer that will be termed spectrometer A. The secondary instruments were a nine-year-old spectrometer (spectrometer B) and a five-year-old instrument (spectrometer C).

The same demountable liquid transmission cell with a 20-mm-diameter circular aperture (model 118-3, Wilmad Glass, Buena, NJ), sapphire windows (Meller Optics, Providence, RI), and a 1.5-mm path length was used to contain the samples in the measurements performed with all three instruments. In previous work, path lengths of 1–2 mm have been used successfully in the same spectral range employed here to quantify analytes in aqueous samples.^{9,10} The sample cell had an integrated water jacket that allowed solution temperatures to be controlled to the physiological range near 37 °C by use of a refrigerated temperature bath. A type T thermocouple was inserted into a port in the sample cell and monitored with a digital thermocouple meter (Omega Engineering, Stamford, CT). Temperature precision was estimated as 37 ± 0.1, 37 ± 0.2, and 37 ± 0.2 °C for measurements made with spectrometers A, B, and C, respectively, on the basis of the standard deviation of the temperatures recorded during the data acquisition. Two similar K-band interference filters, filter 1 and

filter 2 (Barr Associates, Westford, MA), were used to restrict the incident light to the 5000–4000-cm⁻¹ region. Spectrometers A and C used filter 1, and spectrometer B used filter 2. For spectrometer B, a 50% transmittance thin-film-type neutral density filter (Rolyon Optics, Covina, CA) was placed in the optical path to provide a match to the spectral noise levels observed in the data collected with spectrometers A and C.

A pH 6.86, 0.025 M phosphate buffer was prepared with distilled water using 0.025 M KH₂PO₄ and 0.025 M Na₂HPO₄. 5-Fluorouracil (0.1%) was added as a preservative. Every sample was prepared by weighing appropriate amounts of α-D-glucose (ACS reagent, Aldrich Chemical Co., Inc., Milwaukee, WI), sodium L-lactate (99%, Aldrich Chemical Co., Inc.), urea (minimum 99.5%, product no. U-1250, Sigma Chemical Co., St. Louis, MO), sodium L-ascorbate (99+%, Aldrich Chemical Co., Inc.), β-alanine (99+%, Aldrich Chemical Co., Inc.), and triacetin (~99%, Sigma Chemical Co.) and diluting with phosphate buffer to 50 mL. Each sample was split into two approximately equal volumes. The first half was measured using spectrometer A. The second half was frozen and measured 18 days later with spectrometer B. The remaining solution was refrozen and measured 183 days later with spectrometer C.

Procedures. With spectrometer A, data were collected over 7 days during six individual data collection sessions. The data collected with spectrometer B were acquired over 8 days during seven sessions. With spectrometer C, data were collected over 5 days during five sessions. For all the measurements, samples were run in a random order with respect to the component concentrations. The run order was the same for the data collected with each instrument. As a check on sample integrity, glucose concentrations were confirmed with a YSI 2300 STAT PLUS clinical analyzer (YSI, Inc., Yellow Springs, OH) before the near-IR spectra were acquired.

With spectrometers A and C, double-sided interferograms containing 4096 points and based on 256 coadded scans were sampled at every two zero-crossings of the He–Ne laser. With spectrometer B, double-sided interferograms containing 8192 points and based on 256 coadded scans were sampled at every zero-crossing of the laser. For each sample or buffer, three replicate spectra were measured consecutively without removing the sample cell from the spectrometer. Buffer spectra were collected periodically. Interferograms were Fourier processed with triangular apodization, Mertz phase correction, and one level of zero-filling with the manufacturers' software to produce single-beam spectra with a point spacing of 1.9 cm⁻¹. The region of 5000–4000 cm⁻¹ was extracted from all spectra and used for subsequent calculations.

All spectra were transferred to a Silicon Graphics Origin 200 R12000 server (Silicon Graphics, Inc., Mountain View, CA) where subsequent calculations were performed. This computer employed the Irix operating system (Version 6.5, Silicon Graphics, Inc.). By use of original software written in Fortran 77, the single-beam spectra from the three spectrometers were aligned by second-order polynomial interpolation to produce spectra at identical wavenumber points. All further calculations were performed with Matlab (Version 6.0, The MathWorks, Inc., Natick, MA). The PLS_Toolbox 2.1 was used to perform the PDS, DS, and second-derivative calculations (Eigenvector Research, Manson, WA).

(8) Zhang, L.; Small, G. W.; Arnold, M. A. *Anal. Chem.* **2002**, *74*, 4097–4108.

(9) Ding, Q.; Boyd, B. L.; Small, G. W. *Appl. Spectrosc.* **2000**, *54*, 1047–1054.

(10) Pan, S.; Chung, H.; Arnold, M. A.; Small, G. W. *Anal. Chem.* **1996**, *68*, 1124–1135.

Spectra in absorbance units (AU) were obtained by computing ratios of the last replicate spectrum of each sample to the last replicate of the most recently collected buffer sample and converting the resulting transmittance values to absorbance. The other replicate spectra of the samples were not used in this work. The 80 samples were randomly split into a calibration set containing 64 samples and a prediction set with 16 samples.

THEORY

Many multivariate calibration standardization methods have been developed in analytical chemistry.^{11–14} Three methods were compared in this work: DS,⁶ PDS,⁷ and our GMR method.⁸ Each method was used in conjunction with calibration models based on partial least-squares (PLS) regression.^{15,16} A brief description of the DS, PDS, and GMR algorithms is given below.

DS and PDS. Assume that the data measured with both primary and secondary instruments have m spectral channels. The DS method transforms the data from the secondary instrument B to match the primary instrument A by means of a transformation matrix, \mathbf{F} . DS assumes a global linear relationship between the measurement data from instruments A and B:

$$\mathbf{A}_A = \mathbf{A}_B \mathbf{F} + \mathbf{E} \quad (1)$$

where \mathbf{A}_A is an $n \times m$ spectral data matrix for n samples measured with the primary instrument with which the calibration model is built, \mathbf{A}_B is a corresponding $n \times m$ spectral data matrix for these samples measured with the secondary instrument, \mathbf{F} is the $m \times m$ transformation matrix, and \mathbf{E} is the error matrix that contains unmodeled spectral residuals.

The estimated transformation matrix, $\hat{\mathbf{F}}$, can be obtained by measuring a set of k standardization samples with both primary and secondary instruments and using

$$\hat{\mathbf{F}} = \mathbf{A}_{Bs}^+ \mathbf{A}_{As} \quad (2)$$

where \mathbf{A}_{As} and \mathbf{A}_{Bs} are $k \times m$ spectral data matrices for standardization samples obtained with instruments A and B, respectively, and “+” stands for the pseudoinverse. \mathbf{A}_{Bs}^+ can be approximated by singular value decomposition (SVD).¹⁷ A spectrum \mathbf{a}_B measured with the secondary instrument B can be transformed (standardized) to a spectrum \mathbf{a}_{B_std} that pretends to be a spectrum measured with instrument A by using

$$\mathbf{a}_{B_std}^T = \mathbf{a}_B^T \hat{\mathbf{F}} \quad (3)$$

The PDS algorithm is an extension of DS obtained by applying many localized DS models to moving spectral windows.^{6,7} The PDS

method assumes a localized linear relationship between the measurement data for a resolution element from the primary instrument and a measurement window around the corresponding element from the secondary instrument.

Let $\mathbf{a}_{A,i}$ and $\mathbf{a}_{B,i}$ be the i th columns of the matrices \mathbf{A}_A and \mathbf{A}_B , respectively. Assume a spectral window of size $j + k + 1$ has been chosen and that $\mathbf{R}_i = [\mathbf{a}_{B,i-j}, \mathbf{a}_{B,i-j+1}, \dots, \mathbf{a}_{B,i+k}]$ is formed from the spectral data measured with the secondary instrument within the window surrounding resolution element i in the spectra measured with the primary instrument. The PDS algorithm assumes a local model for resolution element i of the primary instrument as follows:

$$\mathbf{a}_{A,i} = \mathbf{R}_i \mathbf{b}_i + \mathbf{e}_i \quad (4)$$

In eq 4, \mathbf{b}_i is a $(j+k+1) \times 1$ regression coefficient vector and \mathbf{e}_i is the model residual vector. By use of PLS or principal component regression (PCR),¹⁶ an estimate of \mathbf{b}_i , $\hat{\mathbf{b}}_i$, can be obtained from the spectra of the standardization samples. By padding zeros at the two ends of $\hat{\mathbf{b}}_i$ corresponding to the resolution elements outside the spectral window and assembling the $\hat{\mathbf{b}}_i$ for each i , a transformation matrix analogous to $\hat{\mathbf{F}}$ in eq 3 can be formed.

When PCR is used to compute the local model, Wise proposed a tolerance value based on the relative sizes of the singular values computed from \mathbf{R}_i as a basis for choosing the number of principal components.¹⁸ A local rank of w will be used if it fulfills

$$\delta_w / \delta_1 > \text{tol} \quad (5)$$

In eq 5, tol is a user-specified threshold tolerance value, δ_1 is the largest singular value, and δ_w is the smallest among the singular values (ordered largest to smallest) that fulfills the above inequality. A default setting of $\text{tol} = 0.01$ was suggested by Wise.

The originally proposed DS and PDS algorithms did not take into consideration the case where a significant baseline difference exists between instruments. This effect can be corrected by removing an additive baseline term.⁷ The removal of the additive term is achieved by mean-centering the spectra of the standardization samples prior to the application of either DS or PDS. This additive background correction procedure was used with DS and PDS in our work.

Selection of Standardization Samples. Leverage calculations were employed to select which samples to use with DS and PDS for computing the transformation matrix and with GMR for updating the calibration model.⁶ Two variations of this calculation are possible. Both use the $n \times m$ mean-centered calibration spectral data matrix \mathbf{A} to compute an $n \times n$ hat matrix, \mathbf{H} . The k th diagonal element of \mathbf{H} is the leverage value of sample k , a measure of the distance of this sample from the center of the data. The spectral hat matrix method defines $\mathbf{H} = \mathbf{A}\mathbf{A}^T$, and the spectral model hat matrix method specifies $\mathbf{H} = \mathbf{A}\mathbf{A}^+$, where \mathbf{A}^+ is the pseudoinverse of \mathbf{A} calculated from the calibration model. The spectral hat matrix method makes selections purely on the basis of the Euclidean distance of each spectrum to the mean, while the spectral model hat matrix approach produces a distance

(11) de Noord, O. E. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 85–87.

(12) Bouveresse, E.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **1996**, *32*, 201–213.

(13) Feudale, R. N.; Woody, N. A.; Tan, H.; Myles, A. J.; Brown, S. D.; Ferre, J. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 181–192.

(14) Heise, H. M.; Winzen, R. In *Near-Infrared Spectroscopy: Principles, Instruments, Applications*; Siesler, H. W., Ozaki, Y., Kawata, S., Heise, H. M., Eds.; Wiley-VCH: Weinheim, 2002; pp 125–162.

(15) Haaland, D. M.; Thomas, E. V. *Anal. Chem.* **1988**, *60*, 1193–1202.

(16) Martens, H.; Næs, T. *Multivariate Calibration*; Wiley: New York, 1989.

(17) Malinowski, E. R. *Factor Analysis in Chemistry*, 2nd ed.; John Wiley & Sons: New York, 1991.

(18) Wise, B. M.; Gallagher, N. B. *PLS Toolbox for use with MATLAB, Version 2.1*; Eigenvector Research, Inc.: Manson, WA, 2000.

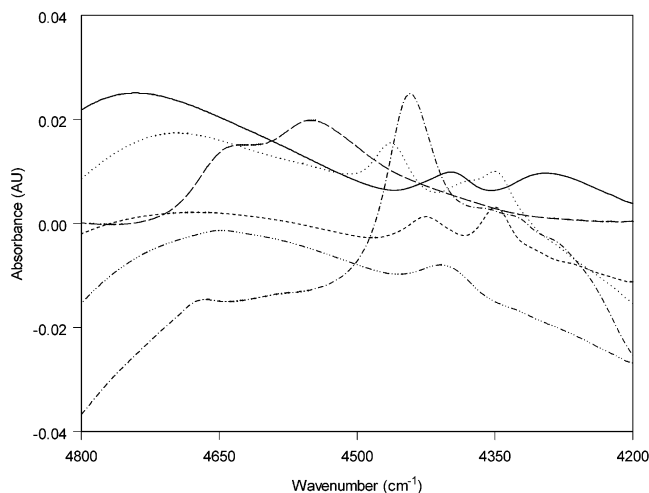


Figure 1. Absorbance spectra of glucose (—), urea (---), triacetin (- · -), ascorbate (·· · · ·), lactate (- - -), and alanine (···). The concentrations were 100 mM for each species except triacetin, whose concentration was 80 mM. A background single-beam spectrum of phosphate buffer was used in the absorbance calculation.

that is weighted by the contribution of sample k to the calibration model. In the work reported here, the spectral hat matrix method was used to select samples for the DS and GMR methods, and the spectral model hat matrix method was used with PDS. The use of different sample selection methods with DS and PDS was adopted on the basis of recommendations in the PLS_Toolbox software documentation.

GMR Algorithm. The GMR algorithm⁸ is a model-updating procedure that uses a set of standardization samples to characterize the response of the secondary instrument and a database of spectra previously acquired with the primary instrument that were used to develop the original calibration model. The spectra of the standardization samples are used to guide an iterative optimization procedure that (1) finds an optimal subset of calibration samples from the original database to use in computing the updated model and (2) finds an optimal set of weights to apply to the spectral resolution elements in order to minimize the effects of instrumental changes on the computed model. The optimization relies on alternating grid search and stepwise sample addition/deletion steps.

RESULTS AND DISCUSSION

Characterization of Spectral Information. Absorbance spectra of individual solutions of alanine, ascorbate, glucose, lactate, triacetin, and urea measured with spectrometer A are presented in Figure 1. These spectra correspond to concentrations (and absorbances) ~ 5 – 7 times higher than the median concentrations used in the mixture samples measured in this research. Extrapolation of the absorbance axis in Figure 1 to the mixture samples reveals that the absorbance signals employed in the calibration work are in the range of 10^{-3} – 10^{-4} AU. Negative absorbance values in the spectra arise from the influence of strong water bands centered near 3800 and 5200 cm^{-1} . The water concentration in the buffer solution used as the reference in the absorbance calculation is higher than in the sample solutions.

In this research, calibration models are built for glucose. Three bands are observed in the glucose spectrum in Figure 1 (solid

line): an O–H combination band centered at 4700 cm^{-1} and two C–H combination bands at 4400 and 4300 cm^{-1} . There is severe overlap between the glucose bands and bands of the other components. For example, previous work has determined that the glucose band at 4400 cm^{-1} is the most significant spectral feature for use in building quantitative models.^{19,20} Alanine, ascorbate, lactate, and triacetin all have C–H combination bands that overlap with this important glucose spectral feature. An effective calibration model for glucose in the 1–30 mM range must reliably extract the 10^{-3} – 10^{-4} AU glucose signals from this variable background. This defines both a challenging calibration problem and an excellent test case for the calibration standardization methodology evaluated in this research.

Characterization of Spectral Quality. To illustrate the differences among the data sets collected with the three instruments, absorbance spectra of four samples measured with spectrometers A, B, and C are displayed in Figure 2. No glucose features are observable in the spectra because of the large background arising from the sample matrix. The spectra were baseline corrected with a two-parameter linear least-squares fit across the 4700–4300- cm^{-1} range. The plots show no large differences among the spectra other than uncorrected baseline variation. This baseline variation arises most likely from instrumental drift and temperature variation between the sample and background spectra used in the absorbance calculation. Other possible contributors are photometric errors and nonlinearities associated with operating the spectrometer at relatively high solution absorbance. At 1.5-mm path length, the absorbance of a buffer sample with respect to the blank cell is 1.40 AU at 4400 cm^{-1} .

During the work with each spectrometer, buffer spectra were collected in groups of three replicates several times during each data collection session. Given that the chemical composition of the buffer is constant, the only variation in these spectra arises from the temperature variation in the sample cell and the short-term and long-term variations in the instrumental response.

To study the differences among these buffer spectra, principal component analysis²¹ (PCA) was performed on each single-beam spectral data set separately using the range of 4700–4300 cm^{-1} . Before the PCA calculation, spectra were mean-centered after normalizing to unit length. Figure 3 displays the first 10 principal component loadings of the buffer spectra collected with spectrometers A, B, and C. In the ideal case, none of these loading spectra would have a nonrandom character. If the instrument were perfectly stable and no temperature variation existed in the sample cell, the mean spectrum would perfectly characterize the data set of buffer spectra, and all loading vectors would consist of random noise.

To identify significant loadings, Shrager and Hendler have reported a method based on the calculation of the first autocorrelation coefficient,²² and Rutledge and Barros have used the Durbin–Watson test for autocorrelation.²³ Both of these approaches are based on the assumption that a nonrandom factor computed from continuous spectra will have correlation between

(19) Arnold, M. A.; Small, G. W. *Anal. Chem.* **1990**, *62*, 1457–1464.

(20) Hazen, K. H.; Arnold, M. A.; Small, G. W. *Appl. Spectrosc.* **1994**, *48*, 477–483.

(21) Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.

(22) Shrager, R. I.; Hendler, R. W. *Anal. Chem.* **1982**, *54*, 1147–1152.

(23) Rutledge, D. N.; Barros, A. S. *Anal. Chim. Acta* **2002**, *454*, 277–295.

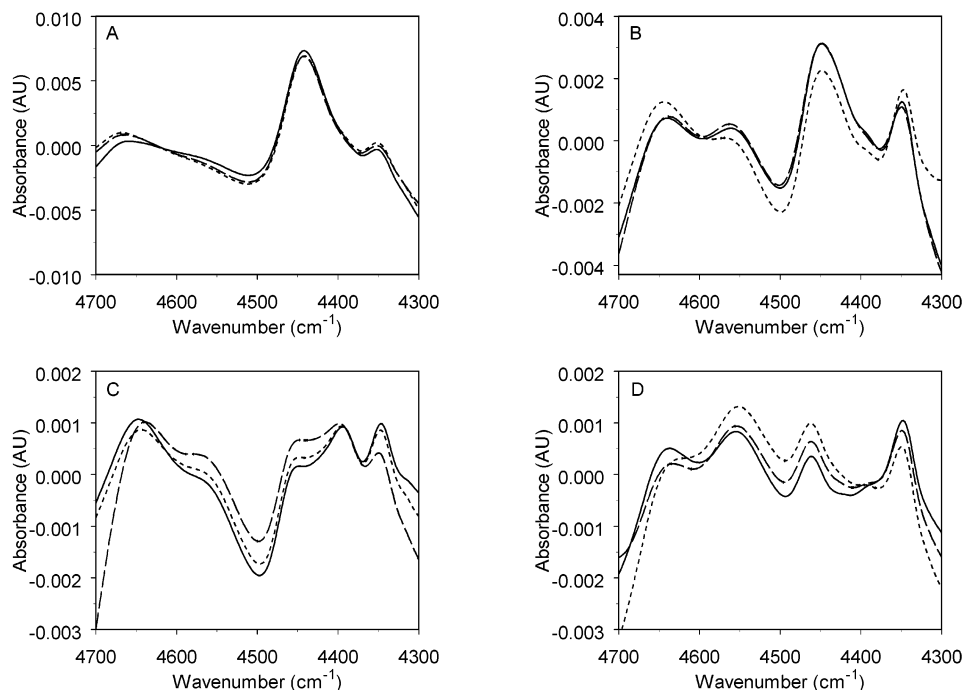


Figure 2. Absorbance spectra of four samples measured with spectrometers A (—), B (---), and C (- - -). Plots A–D correspond to samples 10, 20, 30, and 40, respectively. A spectrum of phosphate buffer was used as the reference in the absorbance calculation. Each spectrum was baseline corrected with a two-parameter linear least-squares fit across the region of 4700–4300 cm^{-1} . No glucose features are observable in these spectra because of the large spectral background arising from the sample matrix.

adjacent points. To apply this methodology to the loadings plotted in Figure 3, the Durbin–Watson test statistic, d , was computed for each factor. Values of d approach zero when the correlation between adjacent points is high.

For all instruments, inspection of Figure 3 reveals the first few factors have similar profiles and are clearly nonrandom. Values of d for factors 1–4 for the three spectrometers are all $< 5.4 \times 10^{-3}$. We hypothesize that these loadings encode the spectral variation associated with the small changes in sample temperature that exist across the data sets. With increasing temperature, the background water absorption changes in magnitude and shifts to higher wavenumber. This effect is nonlinear, and the PCA calculation explains the spectral changes in a piecewise linear manner with several factors.

For spectrometer A, factors 5–10 become progressively more random in character, and the value of d increases in a corresponding manner from 0.014 to 0.103. For spectrometer B, d never exceeds 0.036. These values are in agreement with a visual assessment of Figure 3 that all 10 loadings for spectrometer B have a dominant nonrandom structure. Factor 6 has a harmonic appearance unlike any of the loadings for spectrometers A and C. This suggests that spectrometer B has an instrumental response with quite different variability across the time span of the data. Possible factors contributing to this difference in response could be nonlinearities in the detector and preamplifier or problems in the digital electronics.

For spectrometer C, the values of d for factors 5–10 increase from 5.39×10^{-3} to 0.108. Factors 7–10 all have values of d that exceed the maximum for spectrometer B, while factors 7–9 have smaller d than the corresponding factors from spectrometer A. Overall, visual inspection of Figure 3, coupled with the values of the Durbin–Watson test statistic, suggests that spectrometer A

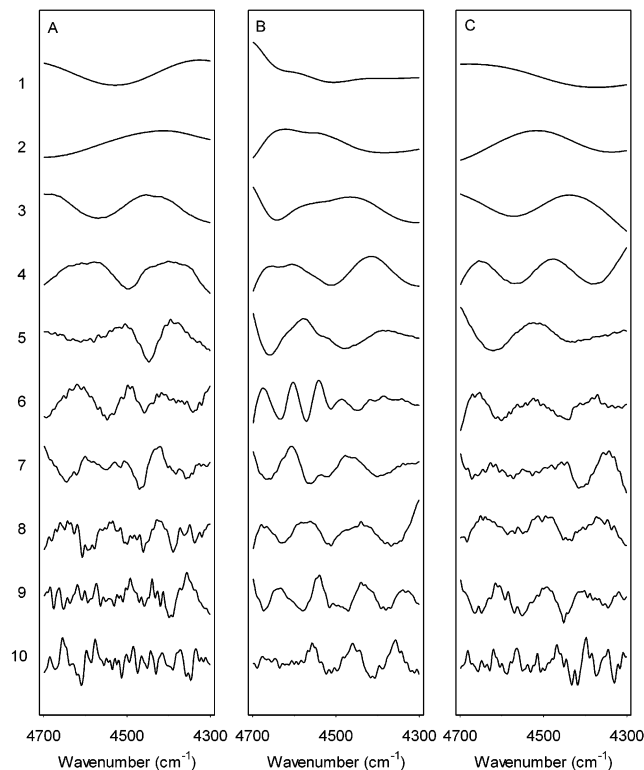


Figure 3. First 10 principal component loadings computed from single-beam spectra of phosphate buffer. (A) Results based on 42 spectra measured with spectrometer A. (B) Results based on 42 spectra measured with spectrometer B. (C) Results based on 45 spectra measured with spectrometer C. The single-beam spectra were normalized to unit length and mean-centered before the PCA calculation.

has the most stable response, followed by spectrometer C. The response of spectrometer B is significantly less stable.

To evaluate the short-term variance of the three data sets, noise levels were estimated for each. Employing the three replicate spectra for each mixture or buffer sample, three noise spectra in AU were obtained by computing the ratios of all combinations of the replicate spectra to each other and converting the resulting transmittance values to absorbance. In the absence of instrumental drift or temperature variation across the replicate spectra, the absorbance noise spectrum will be centered at 0 AU and will be random in character. If the principles of error propagation are applied to the absorbance calculation, the noise value at resolution element i is expected to be proportional to s_i/I_i , where s_i is the noise level in the single-beam spectrum (assumed to be constant with wavenumber for spectra measured with an FT spectrometer) and I_i is the single-beam intensity at i (assumed to be constant across the replicate spectra). The noise in the absorbance spectrum is thus expected to be lowest where the single-beam intensity is highest. A simple change in intensity between the replicate single-beam spectra will cause this noise spectrum to shift above or below 0 AU. Temperature drift between the replicates will cause the noise spectrum to adopt a curved profile arising from a shift in the water absorption bands.

The noise level can be calculated as the root-mean-square (rms) average of the noise values across a selected spectral range. In this study, the region of 4500–4300 cm^{-1} was used because it brackets the important C–H combination region in the vicinity of 4400 cm^{-1} . Three types of noise calculations were performed for comparison purposes. The first, termed $\text{rms}_{\text{total}}$, was a direct calculation of the rms value about an assumed mean of zero. This noise value will be relatively large since it contains all sources of variation. The second procedure, termed rms_{mean} , calculated the noise about the computed mean; i.e., each noise absorbance spectrum was mean-centered by subtracting the mean of the 4500–4300- cm^{-1} region before the rms calculation. This calculation removes simple intensity drift from the noise estimate. The third procedure, termed $\text{rms}_{\text{second-order}}$, calculated the rms noise after a least-squares fit of the noise spectrum to a second-order polynomial function. This calculation removes most of the baseline curvature arising from temperature variation. Except where the temperature variation is large enough such that the second-order polynomial does not fit the curvature closely, the remaining variation is the random component of the noise. These calculations were performed on 94, 94, and 95 samples (including spectra of phosphate buffer) measured with spectrometers A, B, and C, respectively. The corresponding number of noise estimates was 282, 282, and 285.

To assess the short-term variation, the median and interquartile range (IQR) of each group of noise estimates were computed. For $\text{rms}_{\text{total}}$, the medians were 139, 259, and 667 μAU for spectrometers A, B, and C, respectively. The corresponding values of the IQR were 181, 289, and 731 μAU . Median values of rms_{mean} were 21.7, 45.6, and 325 μAU , and the IQR values were 25.7, 45.1, and 380 μAU . For $\text{rms}_{\text{second-order}}$, medians were 4.54, 3.52, and 18.4 μAU . The corresponding IQR values were 2.80, 1.57, and 20.2 μAU .

These results suggest the bulk of the variation between the replicate spectra arises from intensity drift and temperature variation. The data set collected with spectrometer C has by far the largest amount of short-term variation. Visual inspection of the noise spectra in this set revealed curved baselines indicative

of significant temperature variation. This suggests a breakdown in the temperature control procedures used. During the data collection with spectrometer C, problems were encountered in keeping the thermocouple attached to the sample cell. This may have resulted in greater temperature variation and may call the temperatures recorded with the thermocouple into question. The curvature in the absorbance noise spectra in this case was also great enough that it was not completely removed by the second-order polynomial correction used in the $\text{rms}_{\text{second-order}}$ calculation. Thus, these noise estimates are also somewhat affected by the temperature variation.

Overall, spectrometer A exhibits the best short-term performance. Spectrometer B has the lowest random noise (both median value and IQR) but exhibits higher intensity variation than spectrometer A. The temperature variation in this data set also appears to be slightly higher than in the data set collected with spectrometer A.

The study of the long-term and short-term variations indicates that the data set collected with spectrometer A has the highest quality spectra. These data exhibit the least amount of long-term variation, and the short-term variation is better than that of spectrometer B for two of the three measures. The data set collected with spectrometer C has good long-term stability but exhibits significant short-term variation. The data set for spectrometer B has competitive short-term stability, but the long-term variation is by far the worst of the three.

PLS modeling is capable of mitigating the intensity drift, shift in the magnitude and location of the water background absorbance, and random noise that comprise short-term variation. The baseline variation caused by simple linear intensity drift can typically be modeled by one PLS factor. While the effects of temperature variation are nonlinear, they are reproducible and can be modeled in a piecewise linear manner through the introduction of several additional PLS factors. This will be effective as long as the calibration data span the same temperature range and degree of variation that will be encountered when the model is used in prediction. Finally, the uncorrelated nature of random noise will cause the PLS calculation to be biased against it and thus largely force it into the least-significant factors, many of which will likely not be used in building the calibration model.

From the standpoint of its impact on the performance of calibration models, long-term variation is much more problematic than short-term variation. Long-term variation of the type characterized by the loading spectra plotted in Figure 3 is a type of correlated noise. For the calibration model to overcome its effects, additional factors will have to be added. However, these factors are much less stable than, for example, the factors added to model the systematic variation associated with changes in sample temperature. Factors computed to characterize the long-term variation of the calibration data may not extend well to data collected subsequently. Furthermore, for the calibration standardization application that is the focus of this research, the factors that are used to model long-term variation may not extend well across spectrometers.

These considerations suggest that the best-performing calibration models will be obtained with the data collected with spectrometer A, followed by the data measured with spectrometer C. It seems reasonable to project that the long-term variation

evident in the data collected with spectrometer B will cause problems with both calibration performance and calibration standardization.

Spectral Region Selection. Inclusion of spectral channels with poor signal-to-noise (S/N) ratios can degrade the performance of the calibration model significantly.²⁴ Two different region selection methods were used in this study. The first approach employed the region of 4800–4200 cm⁻¹ on the basis of S/N considerations in the absorbance spectra. The second approach sought to optimize the starting and ending spectral points for the specific chemical system used here. A grid search was performed by selecting starting points from the region of 4800–4600 cm⁻¹ and ending points from the range of 4400–4100 cm⁻¹ with an increment of 5 points (9.6 cm⁻¹). Leave-one-block-out cross-validation (CV) was used to choose the best model. The calibration set was randomly split into four blocks. For a given pair of starting and ending spectral points, the following was performed four times. Three blocks of data were used as an internal calibration set for model building. The remaining block was used as an internal prediction set to evaluate the model performance with different numbers of latent variables (LVs) (6–20). On the basis of the CV, a pooled standard error of prediction (CV-SEP) was calculated across the four prediction blocks for a model based on a given number of LVs:

$$CV - SEP = \sqrt{\frac{\sum_{i=1}^n (c_i - \hat{c}_i)^2}{n}} \quad (6)$$

In eq 6, c_i denotes the reference concentration for sample i , \hat{c}_i is the concentration predicted by the model when sample i was withheld from the calibration, and n is the number of samples (spectra) in the calibration set. The CV-SEP corresponding to the minimum among all LVs was chosen for that pair of starting and ending points. Finally, the pair of starting and ending points that gave the minimum CV-SEP was selected. When applied to the calibration data from the primary instrument (spectrometer A), this procedure produced the range of 4760–4242 cm⁻¹.

Parsimonious Model Building. In this study, an F -test at the 95% level was used to reduce the number of LVs and thereby help to avoid overfitting the calibration models.¹⁵ Leave-one-block-out CV as described above was used to choose the number of LVs. After CV, the sum of the squared concentration prediction errors for all calibration samples (termed the prediction error sum of squares or PRESS) was calculated for models with increasing numbers of LVs from 6 to 20. The model selected was the one with the fewest number of factors such that PRESS for that model was not significantly greater than the minimum PRESS.

Evaluation of Performance within the Same Instrument. Calibration models for glucose were built with absorbance data of the 64 calibration samples collected with spectrometers A, B, and C, respectively. The 4800–4200-cm⁻¹ spectral range was used, and the number of LVs in each model was determined by the CV and F -test procedures described above. Each model was evaluated

by application to the 16 prediction samples collected with the same instrument. Figure 4 displays concentration correlation plots for spectrometers A, B, and C, and Table 1 lists the standard error of calibration (SEC), SEP, and prediction bias for each model. Values of SEC are computed analogous to eq 6, with the degrees of freedom adjusted to account for the number of estimated model parameters. Here, SEP is distinguished from CV-SEP to denote that an external prediction set was used with eq 6 rather than a CV calculation. The bias of the prediction data is computed as

$$\text{Bias} = \frac{\sum_{i=1}^n (\hat{c}_i - c_i)}{n} \quad (7)$$

where the terms in eq 7 are as defined previously.

The results presented in Figure 4 and Table 1 show that the calibration and prediction performance of the model for spectrometer A is the best. Spectrometers B and C have similar calibration errors. However, the prediction performance for spectrometer B is the worst and there is significant bias in the prediction. This illustrates the existence of systematic changes in the spectra of prediction set samples that cannot be explained adequately by the calibration model. This result is consistent with the previous discussion regarding the significant long-term variation exhibited by spectrometer B. The overall results are also consistent with our assessment of spectral quality.

Evaluation of Prediction Performance without Standardization. Calibration models for glucose were built with absorbance data from the primary instrument and then applied without the use of any standardization to the spectra of the prediction samples collected with the secondary instruments (spectrometers B and C). Both the 4800–4200- and 4760–4242-cm⁻¹ spectral ranges were used, and the number of factors in each calibration model was determined by the CV and F -test procedures discussed above.

These results showed that the prediction performance of both models within the primary instrument was satisfactory (SEP ≤ 0.53 mM, bias ≤ 0.16 mM, SEC ≤ 0.42 mM). However, both models became invalid when predictions were performed with data from the secondary instruments (SEP ≥ 8.09 mM for spectrometer B and SEP ≥ 3.85 mM for spectrometer C). Optimization of the spectral range improved both the calibration and prediction performances for the primary instrument (SEC reduced from 0.42 to 0.30 mM and SEP reduced from 0.53 to 0.32 mM), but only slightly improved the ability of the model to extend to the data from the secondary instruments (e.g., SEP reduced from 5.34 to 3.85 mM for spectrometer C).

Standard data preprocessing techniques such as second-derivative calculations based on the Savitzky–Golay method^{25,26} were also investigated and were found to have little effect on the calibration and prediction performance with either the primary or secondary instruments.

Comparison of Calibration Standardization Methods. Three calibration standardization methods were compared in this study: DS, PDS, and the GMR algorithm developed in our laboratories. For DS, the number of retained factors used in the

(24) Spiegelman, C. H.; McShane, M. J.; Goetz, M. J.; Motamedi, M.; Yue, Q. L.; Coté, G. L. *Anal. Chem.* **1998**, *70*, 35–44.

(25) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627–1639.

(26) Steinier, J.; Termonia, Y.; Deltour, J. *Anal. Chem.* **1972**, *44*, 1906–1909.

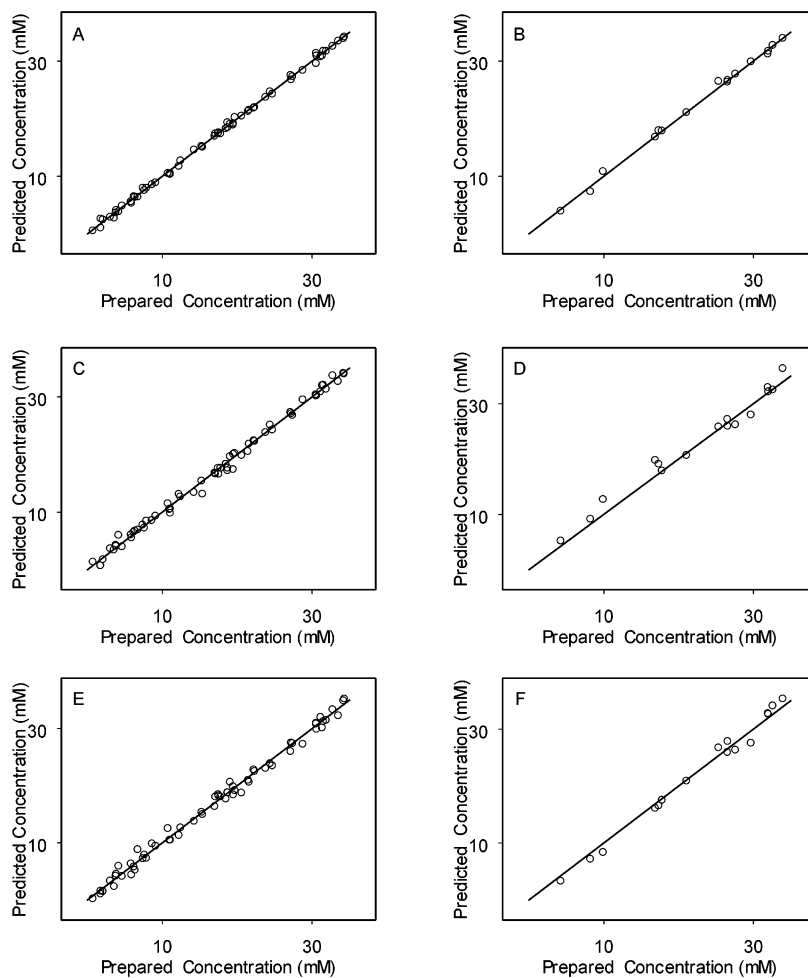


Figure 4. Correlation plots of predicted vs prepared glucose concentrations (mM) for calibration (A, C, E) and prediction samples (B, D, F). Calibration models were developed with the data from each spectrometer and then applied to the spectra of the prediction samples collected with the same instrument. Results are plotted for spectrometers A (plots A and B), B (plots C and D), and C (plots E and F). Absorbance spectra in the range of 4800–4200 cm^{-1} were used in model building. The solid line in each plot denotes perfect correlation.

Table 1. Model Performance for Glucose within the Same Instrument^a

spectrometer	no. of LVs	calibration		prediction	
		SEC (mM)	SEP (mM)	SEP (mM)	bias (mM)
A	12	0.42	0.53	0.53	0.16
B	12	0.80	1.51	1.51	0.72
C	11	0.95	1.14	1.14	-0.05

^a Calibration models computed with the spectral region of 4800–4200 cm^{-1} .

SVD approximation of A_{Bs}^+ was chosen according to the default setting in the PLS_TOOLBOX software, i.e., one less than the number of standardization samples. For PDS, two parameters need to be determined, i.e., the tolerance value and the window size. They were optimized by a grid search. The tolerance value was chosen to fall within the following intervals: 1.0×10^{-5} – 1.0×10^{-4} , 1.0×10^{-4} – 1.0×10^{-3} , 1.0×10^{-3} – 1.0×10^{-2} , and 1.0×10^{-2} – 1.0×10^{-1} . Each interval was evenly split into 10 levels. The window size was chosen to lie between 3 and 21 data points with an increment of 2. The combinations of window size and tolerance value that gave the minimum SEP for the standardization samples

were chosen. During this work, it was observed that, in many cases, the optimized window size hit the boundary of 21 points. Studies were then performed in a stepwise manner to increase the upper limit of the window size. Even with a window size of 81, the optimization hit the window size limit in some cases. The results showed, however, that there is no significant improvement in the SEP for the prediction samples when a larger window size is used. Accordingly, the results presented here reflect a maximum window size of 21 points.

In initial studies, the performance of PDS was by far the worst among the three algorithms. The SEP values exceeded those obtained without standardization, even after using 17 standardization samples. Figure 5A displays PDS-transformed absorbance spectra for the prediction set when three standardization samples were used. These spectra exhibit pronounced discontinuities. This characteristic of PDS has been reported previously by Gemperline et al.²⁷ This work also reported that optimization of the window size and number of factors used with PDS cannot guarantee the elimination of these discontinuities.

(27) Gemperline, P. J.; Cho, J.-H.; Aldridge, P. K.; Sekulic, S. S. *Anal. Chem.* **1996**, *68*, 2913–2915.

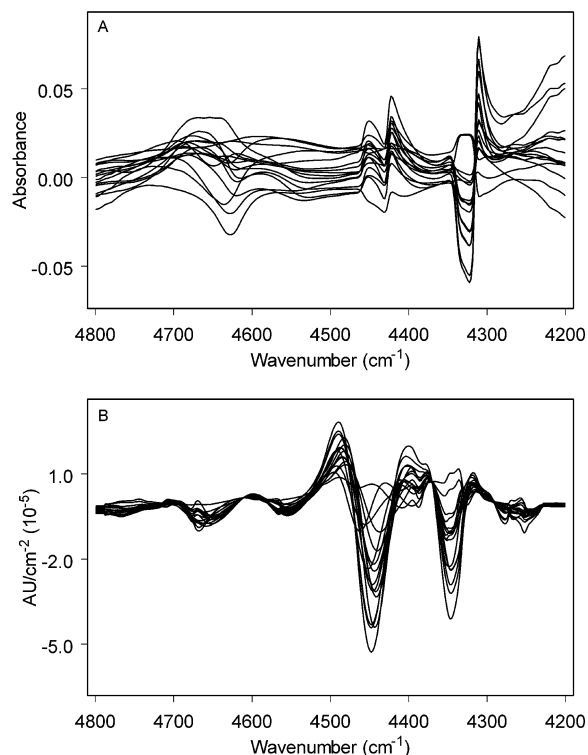


Figure 5. PDS-transformed absorbance (A) and second-derivative (B) spectra of the prediction set for spectrometer C. For both absorbance and derivative spectra, the optimized PDS window size was 21 points and the tolerance value was 1×10^{-5} . Discontinuities are readily apparent in plot A.

Figure 5B displays the transformed data after second-derivative spectra were submitted to the PDS algorithm. The discontinuity artifacts are less apparent in these spectra, and it was discovered that the SEP values obtained with PDS-transformed spectra were much better when second-derivative preprocessing was performed. For this reason, all results presented here for the PDS algorithm are based on second-derivative spectra. The use of second-derivative preprocessing did not improve the results obtained with either the DS or GMR algorithms. Accordingly, the raw absorbance data were used as the inputs for DS and GMR.

The use of second-derivative data with the PDS calculations required optimization of the derivative computation. A grid search combined with CV was performed with the calibration data of the primary instrument to find the optimal combination of the window size and the order of polynomial fitting associated with the Savitzky–Golay polynomial derivatives employed here.^{25,26} Window sizes ranged from 7 to 31 with an increment of 2, and polynomial fitting was tested with second and third orders. Leave-one-block-out CV was used to choose the best model as described previously. Values of CV-SEP were calculated with LVs ranging from 6 to 20. The combination of parameters that gave the minimum CV-SEP was chosen. For the 4800–4200- and 4760–4242-cm⁻¹ ranges, respectively, this optimization selected (order 2, window size 29) and (order 3, window size 31). These values were used throughout the work with PDS for the respective spectral ranges.

For the GMR algorithm, a grid search based on 11 levels was used in weight determination, a value of 0.05 was employed in the initial deletion process based on PRESS values, and models were restricted to a maximum of 12 LVs.⁸ *F*-Tests used internally

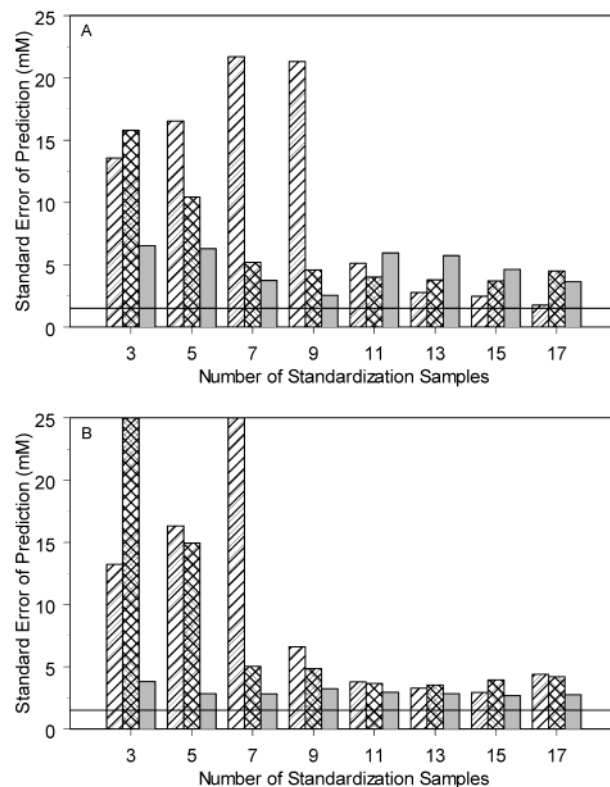


Figure 6. Clustered bar graphs comparing SEP values for the prediction samples when calibration standardization was applied to the data from spectrometer B. Spectrometer A served as the primary instrument. Each group of bars presents results for the DS (striped bars), PDS (crosshatched bars), and GMR (solid bars) algorithms. A horizontal line is drawn at SEP = 1.51 mM corresponding to the prediction performance obtained when a model was developed with the calibration data from spectrometer B (see Table 1). Plots A and B present results from the 4800–4200- and 4760–4242-cm⁻¹ ranges, respectively. None of the results meet the desired level of performance.

for the selection of the number of LVs and after algorithm completion for determining the final model for use in prediction were performed at the 70% level.

As noted previously, the DS and GMR algorithms used the spectral hat matrix method to select standardization samples, while PDS used the spectral model hat matrix method. Note that although the GMR method does not require the same sets of samples to be measured with both primary and secondary instruments, the same standardization samples employed with DS were used with the GMR method to facilitate comparisons.

Figure 6 displays bar graphs that summarize the SEP values obtained with the prediction set for spectrometer B as a result of employing the three calibration standardization methods. Separate graphs are plotted for the 4800–4200- (Figure 6A) and 4760–4242-cm⁻¹ spectral ranges (Figure 6B). Striped, crosshatched, and solid bars correspond to the use of the DS, PDS, and GMR algorithms, respectively, and results are presented across the range of 3–17 standardization samples. A horizontal line is drawn in each plot at 1.51 mM corresponding to the SEP value obtained when the data from spectrometer B were used internally to build the calibration model (Table 1). This SEP value represents the target level of performance for the calibration standardization algorithms when applied to spectrometer B.

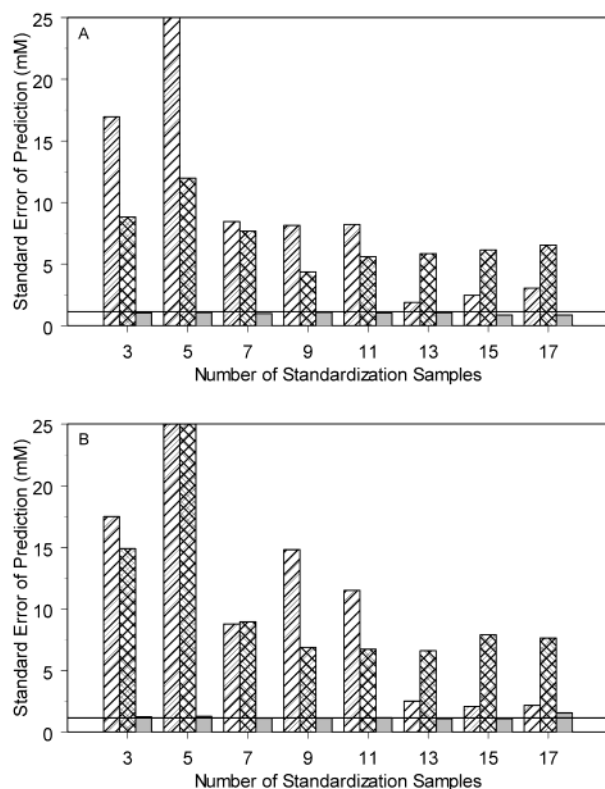


Figure 7. Clustered bar graphs comparing SEP values for the prediction samples when calibration standardization was applied to the data from spectrometer C. Spectrometer A served as the primary instrument. Each group of bars presents results for the DS (striped bars), PDS (crosshatched bars), and GMR (solid bars) algorithms. A horizontal line is drawn at SEP = 1.14 mM corresponding to the prediction performance obtained when a model was developed with the calibration data from spectrometer C (see Table 1). Plots A and B present results from the 4800–4200- and 4760–4242-cm⁻¹ ranges, respectively. Only the GMR algorithm is able to match the desired level of performance.

Inspection of Figure 6 reveals that none of the algorithms performs satisfactorily in effecting calibration transfer to spectrometer B. The GMR results are the most stable with respect to the number of standardization samples, but all of the results exceed the desired SEP value. The DS and PDS algorithms require ~11 standardization samples to achieve a stable result. All of the algorithms realize some benefit from the use of the optimized 4760–4242-cm⁻¹ spectral range, but the overall results are still unsatisfactory. The variation in the data from spectrometer B discussed previously appears to limit the applicability of calibration transfer from spectrometer A to spectrometer B.

Figure 7 presents the analogous results for the application of calibration transfer from spectrometer A to spectrometer C. The format of the figure is identical to Figure 6. The target level of performance is again indicated by the horizontal line drawn at SEP = 1.14 mM, corresponding to the internal prediction result when the data from spectrometer C were used to build the calibration model.

An inspection of Figure 7 reveals that the GMR algorithm performs very well. The results are stable with respect to the number of calibration samples, and the target level of performance is achieved. The PDS algorithm performs poorly, even with 17 standardization samples. The DS algorithm stabilizes after 13

standardization samples, but the SEP values obtained never reach the desired performance level. There is little difference in the results between the two spectral ranges tested.

The results presented in Figures 6 and 7 illustrate that successful calibration standardization requires a certain level of compatibility between the data collected with the primary and secondary spectrometers. The presence of significantly different sources of variation between the primary and secondary instruments (e.g., the case of spectrometers A and B here) can prevent an effective calibration standardization. This factor emphasizes the importance of diagnostic tests for assessing the potential success of calibration transfer.

For the GMR algorithm, we have developed such a diagnostic procedure based on varying the number of standardization samples and evaluating both the number of LVs in the updated calibration model and the SEP for the standardization samples (SEP_{xfer}). For calibration transfer to spectrometer B, Figure 8A presents plots of SEP_{xfer} with respect to the number of standardization samples. Figure 8B is the corresponding plot for spectrometer C. Horizontal lines are included in the figures to indicate the SEP values that define the desired level of performance of each instrument (Table 1). Panels C and D of Figure 8 are the corresponding plots of the number of LVs in the updated model as a function of the number of standardization samples for calibration transfer to spectrometers B and C, respectively. Horizontal lines are drawn at LV = 12 in these plots corresponding to the original model size optimized for the calibration data (spectrometer A). In all the plots, results are plotted for both the 4800–4200- (circles) and 4760–4242 (triangles)-cm⁻¹ spectral ranges.

For a successful calibration transfer, the expectation is that SEP_{xfer} will increase and then stabilize as the number of standardization samples is increased. Once stable, the value of SEP_{xfer} should define a best-case estimate for SEP values to be obtained subsequently when the model is applied to data collected with the secondary instrument. Corresponding plots of the number of LVs should also stabilize and reach a value close to that optimized previously with the primary instrument.

Inspection of Figure 8 reveals that the expected behavior described above is achieved for spectrometer C but not for spectrometer B. For spectrometer C, SEP_{xfer} stabilizes at nine standardization samples and the plot trace suggests the desired level of performance of SEP = 1.14 mM should be attainable with models based on either spectral range. As expected, the updated number of LVs is close to the original value of 12. By contrast, SEP_{xfer} for spectrometer B never reaches the performance level expected. For the 4800–4200-cm⁻¹ range, the trace also never stabilizes. In addition, the number of LVs is much lower than expected. For the 4800–4200-cm⁻¹ range, this trace is also unstable. Panels A and C of Figure 8 suggest the model based on 4760–4242 cm⁻¹ should be the better of the two, but the level of performance is estimated to be unsatisfactory. When the results presented in Figures 6–8 are considered together, the diagnostics plotted in Figure 8 are clearly effective in estimating the performance of the calibration transfer with the prediction samples.

As a final assessment of the GMR method, models were selected corresponding to the smallest number of standardization samples that produced stable traces in Figure 8 and applied to

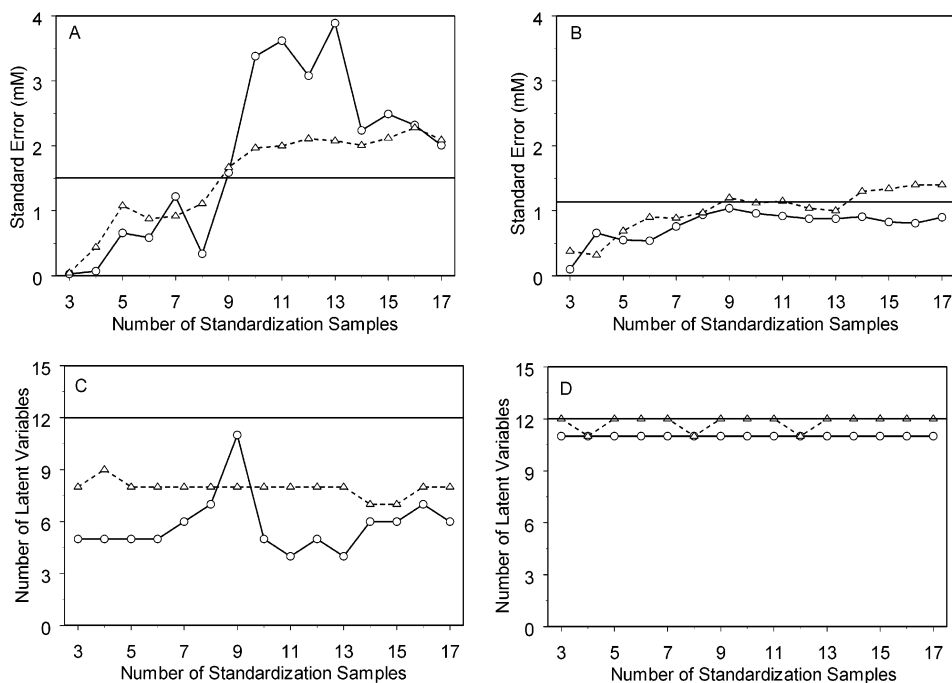


Figure 8. Diagnostics for predicting the performance of the GMR algorithm on the basis of the results obtained with the standardization samples. Values of SEP_{xfer} are plotted as a function of the number of standardization samples for the calibration transfer from spectrometer A to spectrometer B (plot A) and spectrometer C (plot B). The horizontal line in each plot indicates the target performance level corresponding to the SEP value obtained when the data from the secondary instrument were used to build the calibration model (see Table 1). Plots of the number of LVs in the updated calibration models are also provided as a function of the number of standardization samples for spectrometers B (plot C) and C (plot D). The horizontal line in each plot denotes $LV = 12$, the size of the initial calibration model built with the data from the primary instrument (spectrometer A). In all plots, results are presented for both the 4800–4200- (circles) and 4760–4242 (triangles)- cm^{-1} spectral regions.

the independent prediction sets from spectrometers B and C. For spectrometer B, the model based on 11 standardization samples and the 4760–4242- cm^{-1} range produced values of $SEP_{xfer} = 2.00$ mM and $SEP = 2.96$ mM. For spectrometer C, values of $SEP_{xfer} = 1.04$ mM and $SEP = 1.07$ mM were obtained with the model based on nine standardization samples and the 4800–4200- cm^{-1} range. Comparison of these values to the within-instrument prediction results listed in Table 1 confirms that calibration transfer is very successful with spectrometer C but that the prediction results for spectrometer B are degraded when calibration standardization is performed. Concentration correlation plots similar to those in Figure 4 visually confirmed these conclusions. For spectrometer C, the plot appeared virtually identical to Figure 4F. The analogous plot for spectrometer B was significantly more scattered than Figure 4D.

CONCLUSIONS

For the challenging near-IR glucose determination discussed here, the GMR algorithm significantly outperformed DS and PDS. For calibration transfer between spectrometers A and C, the GMR procedure provided stable performance with as few as three standardization samples. The availability of procedures for predicting the success of calibration standardization is an important

characteristic of the GMR method. These diagnostics correctly forecasted the successful calibration transfer with spectrometer C and the unsatisfactory results obtained with spectrometer B.

The disparity in results obtained with spectrometers B and C underscores the point that successful calibration standardization requires a match in the variance structures associated with the data from the primary and secondary spectrometers. As displayed in Figure 3, the presence of mismatching principal components in the phosphate buffer data is a clear indicator of potential problems in attempting to perform calibration standardization. This factor is particularly important in applications such as the one discussed here where the calibration is based on the extraction of very small analyte signals from a highly variable background.

ACKNOWLEDGMENT

Jun Chen is acknowledged for his help in collecting the data used in this research. Funding for this work was provided by the National Institutes of Health under Grant DK60657.

Received for review May 9, 2003. Accepted August 21, 2003.

AC034495X